DATA ANONYMIZATION EXERCISE - 2023

BUSINESS AND PRIVACY ISSUES IN DATA ANALYTICS

BUSINESS ANALYTICS

DEADLINE: 4/4/2023

MANOLIS TERROVITIS

The US Census Bureau provides microdata to be used in research and applications. These data were deemed a threat to use privacy and in 2020 they were anonymized for the first time. In this exercise you must study the last non-anonymized census and discuss the privacy threats. An example of the microdata of the 2010 census can be found here: https://www.census.gov/data/datasets/2010/dec/stateside-pums.html

For a full description of the dataset you can read https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/us-pums/pumsus.pdf

For the exercise, please use the Delaware data: https://www2.census.gov/census_2010/12-Stateside_PUMS/Delaware/

How to use the dataset

In order to use the dataset, download and decompress the zip file.

The extracted directory contains the de.2010.pums.01.txt

The de.2010.pums.01.txt file contains rows corresponding to:

(a) Person records rows start with ‘P’
(b) Household records rows that start with ‘H’

For the next steps ignore the Household records.

For the Person records please read the brief description for the comprised columns, and the possible values for each column that can be found in:

https://www2.census.gov/census_2010/12-Stateside_PUMS/2010%20PUMS%20Record%20Layout.xlsx

We suggest using software like Excel or Numbers (manually) to guide the Person data or write a script to parse the file into a table.

Exercise A

After examining the data table (columns & values), answer the following questions:

1. Which attributes can act as quasi-identifiers and why?
2. Which of the following properties holds for the data?
    a. They are anonymized
    b. They are pseudonymized
    c. They are encrypted

    Explain the key differences between the three approaches with respect to GDPR.
3. Explain how a person can be identified.
4. Define differential privacy and explain the importance of the privacy parameter e.

Exercise B

Load the dataset into a Python notebook (we suggest Jupyter) and display the first few rows to understand the data.

1. Use the Amnesia anonymization tool to apply k-anonymity to the dataset. Comment on the resulting dataset.
2. Plot the distribution of numeric features in the dataset using histograms.
3. Apply a random noise mechanism to some of the numeric columns using the Gaussian mechanism. The noise should be added to the original values in a way that preserves differential privacy.
4. Calculate the differentially private averages for the individuals using the noisy data.
5. Plot the distribution of numeric features after the noise addition. Try different values of the e parameter. Comment on the effect of the differential privacy on the results.

* For steps 2-5 you can use the https://github.com/IBM/differential-privacy-library and https://github.com/IBM/differential-privacy-library/tree/main/notebooks.